

Contents

1	Introduction	3
1.1	Context and Importance	3
2	Project Objectives	4
3	Azure Setup and User Management	5
3.1	User Creation in Azure Entra ID	5
3.2	Resource Setup	6
3.3	SCIM Integration with Databricks	6
4	Data Architecture	7
4.1	Azure Data Lake Storage	7
4.2	Bronze Layer	8
4.3	Silver Layer	9
4.4	Gold Layer	9
5	Role-Based Access Control and Row-Level Security	10
5.1	RBAC with Unity Catalog	10
5.2	Row-Level Security	12
6	Dashboards	14
7	Data Cataloging with Unity Catalog	15
7.1	Centralized Metadata Repository	15
7.2	Enhanced Data Governance	16
8	Lessons Learned	17
9	Conclusion	18
10	Code File	19
	References	20

List of Figures

3.1	User Accounts Created in Azure Entra ID	5
3.2	Resource Groups and Storage Accounts	6
3.3	SCIM Integration for Databricks	6
4.1	External Storage connection created in Databricks	7
4.2	Raw Data files present in ADLS	8
4.3	Bronze Layer Architecture	8
4.4	Silver Layer Architecture	9
4.5	Gold Layer Architecture	9
5.1	Role-Based Access for Sales User	11
6.1	Role-based dashboards with restricted access.	14
7.1	Data Lineage in Unity Catalog	15

Chapter 1

Introduction

This report outlines a comprehensive implementation of a Data Access Management System for analytics platform. The goal is to design a robust, secure, and scalable architecture that enables role-based access to critical data assets using Azure Databricks, Unity Catalog, and Azure Data Lake Storage.

1.1 Context and Importance

The project aims to simulate an enterprise-level data access management process to enhance data governance, ensuring sensitive data is accessed securely while empowering stakeholders with actionable insights. Tools such as Azure Entra ID, Databricks SCIM, and Unity Catalog were employed to achieve these objectives.

Chapter 2

Project Objectives

The following objectives guided the implementation:

- Implement a secure data platform using Microsoft Azure and Databricks.
- Establish a multi-layered data architecture (Bronze, Silver, Gold).
- Apply Role-Based Access Control (RBAC) and Row-Level Security (RLS).
- Develop dashboards for actionable insights with role-based access.
- Enable data cataloging for enhanced metadata management and data lineage tracking using Unity Catalog.

Chapter 3

Azure Setup and User Management

3.1 User Creation in Azure Entra ID

The following users were created in Azure Entra ID to simulate real-world roles:

- **data_engineer1@mukeshbackup09gmail.onmicrosoft.com** - Responsible for managing data pipelines and processing.
- **analyst1@mukeshbackup09gmail.onmicrosoft.com** - Germany-based analyst focusing on regional data insights.
- **business_analyst1@mukeshbackup09gmail.onmicrosoft.com** - Business analyst with access to all high-level metrics.
- **sales_user1@mukeshbackup09gmail.onmicrosoft.com** - Focused on sales data for specific territories.
- **marketing_user1@mukeshbackup09gmail.onmicrosoft.com** - Marketing specialist analyzing product performance and customer behavior.

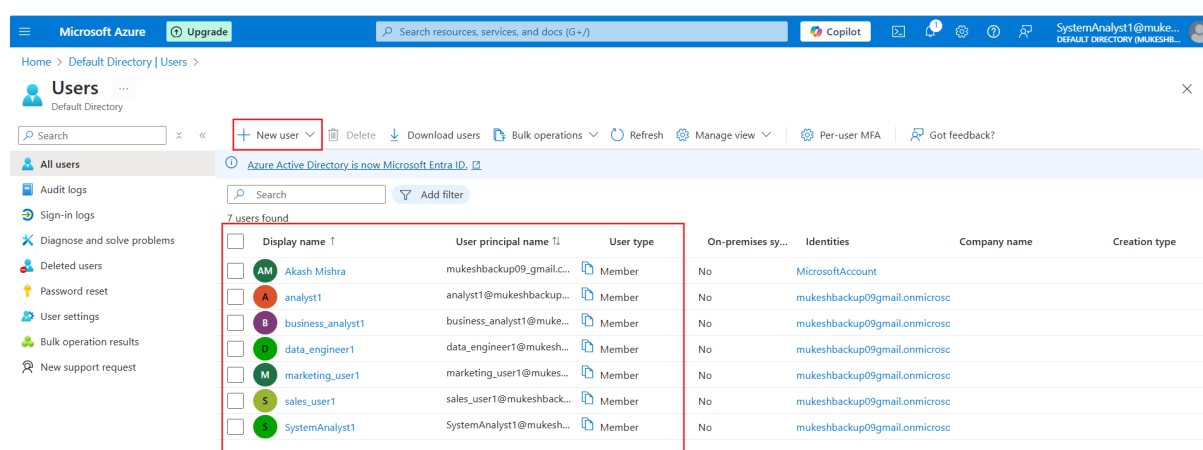


Figure 3.1: User Accounts Created in Azure Entra ID

3.2 Resource Setup

Creating resource groups, connectors, storage accounts, and the Databricks workspace.

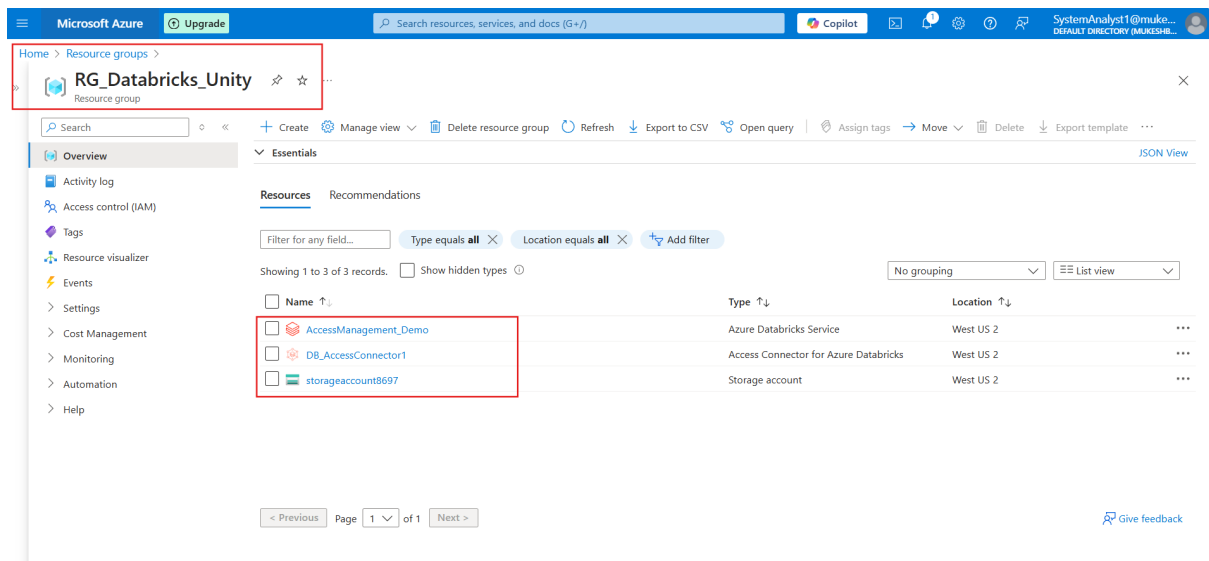


Figure 3.2: Resource Groups and Storage Accounts

3.3 SCIM Integration with Databricks

To automate user provisioning and synchronize roles between Azure Entra ID and Databricks, SCIM (System for Cross-domain Identity Management) was configured.

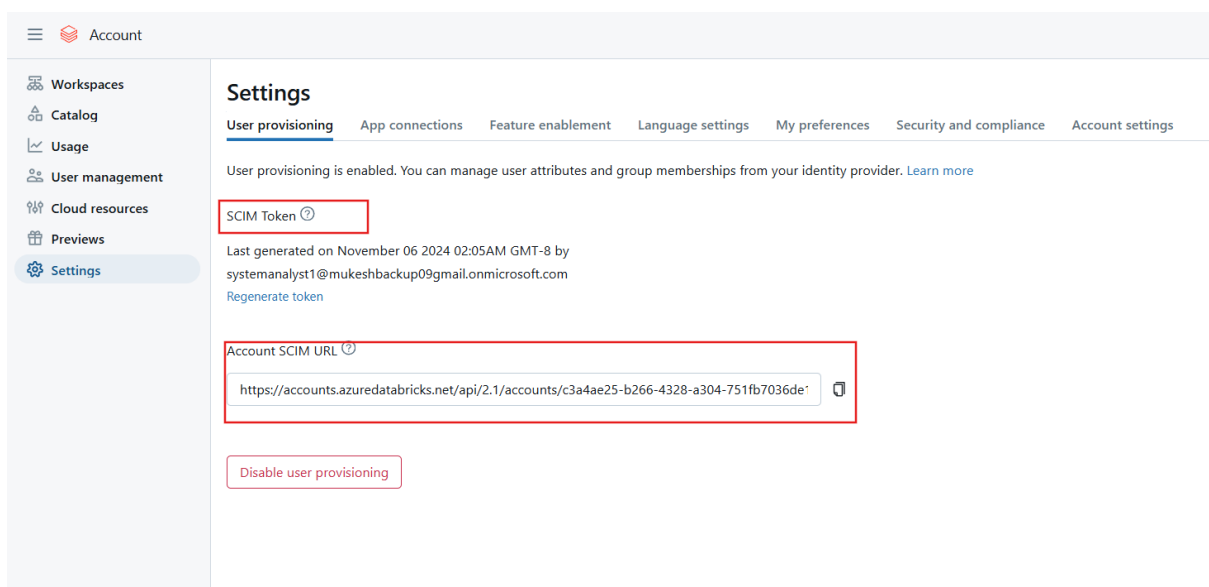


Figure 3.3: SCIM Integration for Databricks

Chapter 4

Data Architecture

4.1 Azure Data Lake Storage

Raw CSV files of Adventure Works Company have been placed in ADLS storage and a storage connection has been built with Databricks.

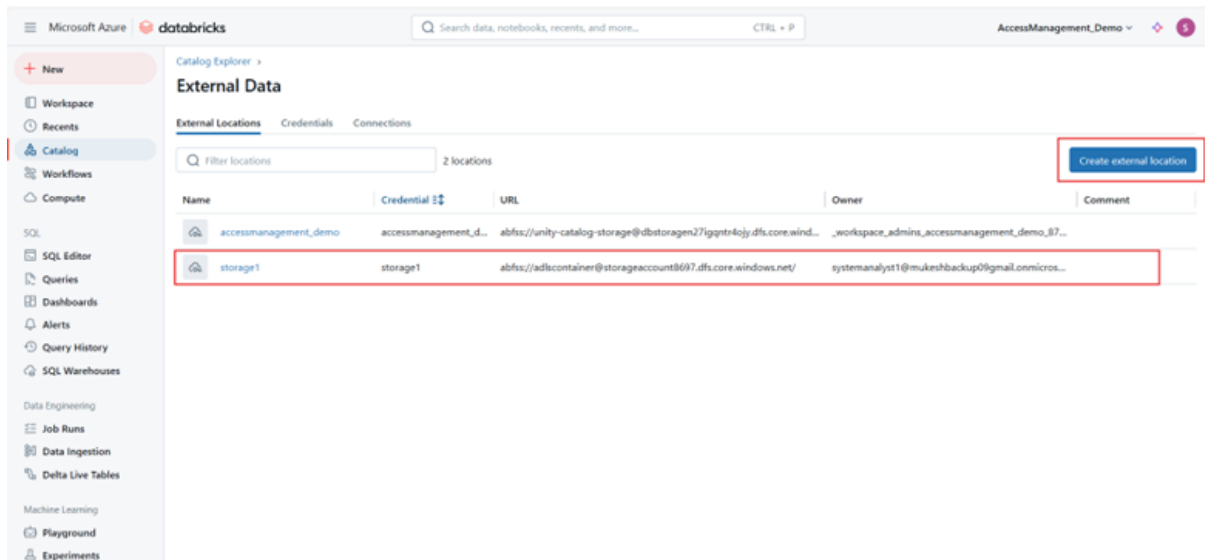


Figure 4.1: External Storage connection created in Databricks

Name	Modified	Access tier	Archive status	Blob type	Size	Lease state
AdventureWorks Calendar Lookup.csv	11/6/2024, 12:36:48 ...	Hot (Inferred)		Block blob	10.69 KiB	Available
AdventureWorks Customer Lookup.csv	11/6/2024, 12:36:52 ...	Hot (Inferred)		Block blob	1.8 MiB	Available
AdventureWorks Product Categories Lookup.csv	11/6/2024, 12:36:49 ...	Hot (Inferred)		Block blob	83 B	Available
AdventureWorks Product Lookup.csv	11/6/2024, 12:36:49 ...	Hot (Inferred)		Block blob	56.76 KiB	Available
AdventureWorks Product Subcategories Lookup.c...	11/6/2024, 12:36:49 ...	Hot (Inferred)		Block blob	637 B	Available
AdventureWorks Returns Data.csv	11/6/2024, 12:36:49 ...	Hot (Inferred)		Block blob	35.58 KiB	Available
AdventureWorks Sales Data 2020.csv	11/6/2024, 12:36:49 ...	Hot (Inferred)		Block blob	121.06 KiB	Available
AdventureWorks Sales Data 2021.csv	11/6/2024, 12:36:52 ...	Hot (Inferred)		Block blob	1.08 MiB	Available
AdventureWorks Sales Data 2022.csv	11/6/2024, 12:36:53 ...	Hot (Inferred)		Block blob	1.32 MiB	Available
AdventureWorks Territory Lookup.csv	11/6/2024, 12:36:49 ...	Hot (Inferred)		Block blob	400 B	Available
Product Category Sales (Unpivot Demo).csv	11/6/2024, 12:36:50 ...	Hot (Inferred)		Block blob	632 B	Available

Figure 4.2: Raw Data files present in ADLS

4.2 Bronze Layer

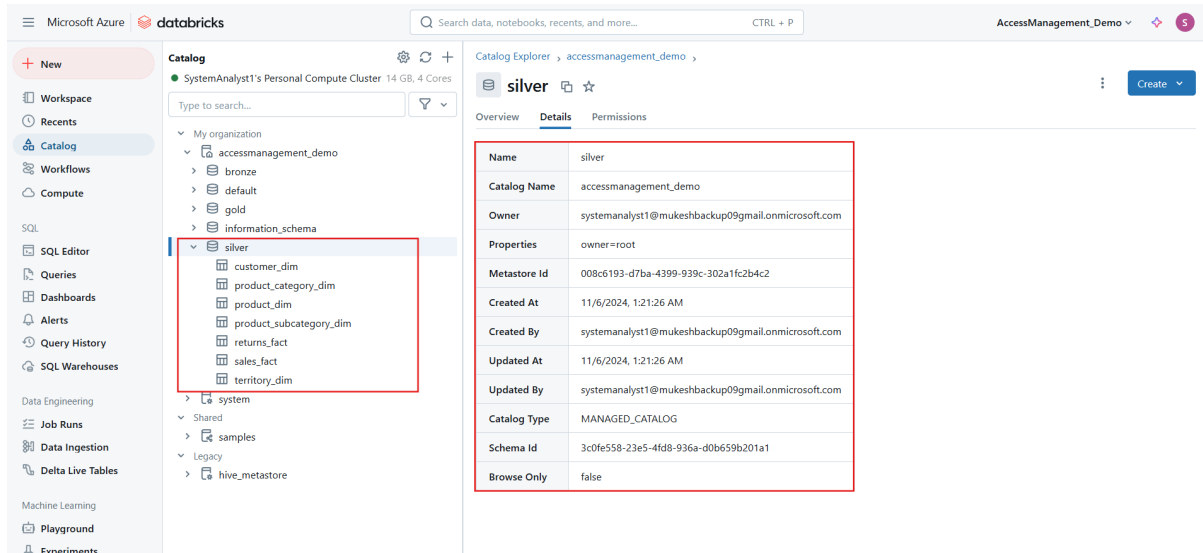
Raw data ingested from the Azure Data Lake Storage (ADLS) container into tables. This layer stores untransformed data.

Name	Owner	Created at	Popularity
customer	systemanalyst1@mukeshback...	2024-11-06 01:22:49	
product	systemanalyst1@mukeshback...	2024-11-06 01:22:55	
product_categories	systemanalyst1@mukeshback...	2024-11-06 01:23:00	
product_subcategories	systemanalyst1@mukeshback...	2024-11-06 01:23:05	
returns	systemanalyst1@mukeshback...	2024-11-06 01:23:14	

Figure 4.3: Bronze Layer Architecture

4.3 Silver Layer

Cleaned and standardized data for analytical use. This layer addresses data quality and structure improvements.



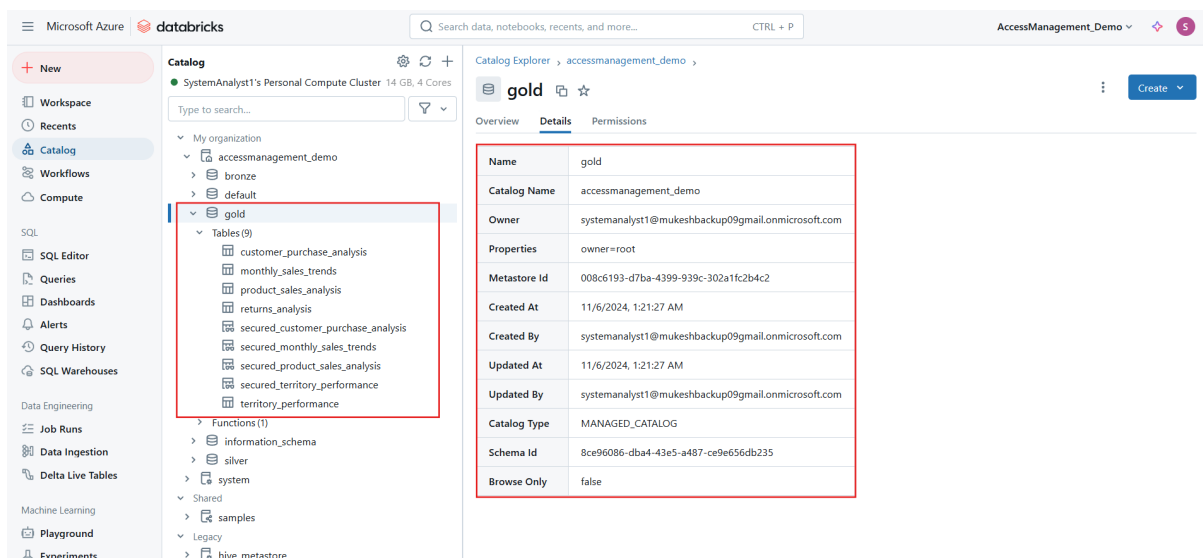
The screenshot shows the Databricks Catalog Explorer interface. On the left, the 'Catalog' tree is expanded to show the 'silver' layer, which contains several tables: customer_dim, product_category_dim, product_dim, product_subcategory_dim, returns_fact, sales_fact, and territory_dim. On the right, the 'Details' tab for the 'silver' catalog is displayed, showing the following information:

Name	silver
Catalog Name	accessmanagement_demo
Owner	systemanalyst1@mukeshbackup09gmail.onmicrosoft.com
Properties	owner=root
Metastore Id	008c6193-d7ba-4399-939c-302a1fc2b4c2
Created At	11/6/2024, 1:21:26 AM
Created By	systemanalyst1@mukeshbackup09gmail.onmicrosoft.com
Updated At	11/6/2024, 1:21:26 AM
Updated By	systemanalyst1@mukeshbackup09gmail.onmicrosoft.com
Catalog Type	MANAGED_CATALOG
Schema Id	3c0fe558-23e5-4fd8-936a-d0b659b201a1
Browse Only	false

Figure 4.4: Silver Layer Architecture

4.4 Gold Layer

Aggregated, business-ready data for use in dashboards and reporting.



The screenshot shows the Databricks Catalog Explorer interface. On the left, the 'Catalog' tree is expanded to show the 'gold' layer, which contains several tables: customer_purchase_analysis, monthly_sales_trends, product_sales_analysis, returns_analysis, secured_customer_purchase_analysis, secured_monthly_sales_trends, secured_product_sales_analysis, secured_territory_performance, and territory_performance. On the right, the 'Details' tab for the 'gold' catalog is displayed, showing the following information:

Name	gold
Catalog Name	accessmanagement_demo
Owner	systemanalyst1@mukeshbackup09gmail.onmicrosoft.com
Properties	owner=root
Metastore Id	008c6193-d7ba-4399-939c-302a1fc2b4c2
Created At	11/6/2024, 1:21:27 AM
Created By	systemanalyst1@mukeshbackup09gmail.onmicrosoft.com
Updated At	11/6/2024, 1:21:27 AM
Updated By	systemanalyst1@mukeshbackup09gmail.onmicrosoft.com
Catalog Type	MANAGED_CATALOG
Schema Id	8ce96086-dba4-43e5-a487-ce9e656db235
Browse Only	false

Figure 4.5: Gold Layer Architecture

Chapter 5

Role-Based Access Control and Row-Level Security

5.1 RBAC with Unity Catalog

Unity Catalog was used to enforce fine-grained RBAC. The following role definitions and permissions were implemented:

Data Engineers

- Full access to all layers (Bronze, Silver, Gold)
- Can create, modify, and delete tables
- Responsible for ETL processes and data pipeline maintenance

Data Analysts

- Read access to Silver Layer
- Full access to Gold Layer
- Can create and modify analytical views
- Can perform ad-hoc analysis

Business Analysts

- Read-only access to Gold Layer
- Can view all business-ready data
- Access to all analytical views

Sales Team

- Access to specific Gold Layer tables:
 - Product sales analysis
 - Customer purchase analysis
 - Monthly sales trends
 - Territory performance
- Row-level security based on region

Marketing Team

- Access to specific Gold Layer tables:
 - Customer purchase analysis
 - Product sales analysis
 - Returns analysis

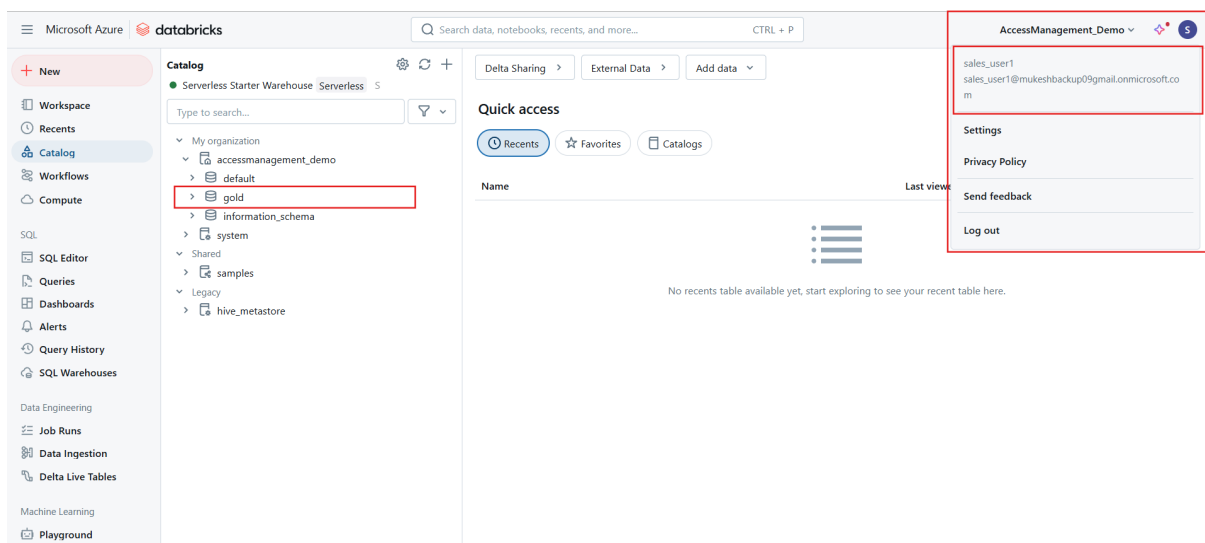


Figure 5.1: Role-Based Access for Sales User

5.2 Row-Level Security

1.1 RLS View for gold.product_sales_analysis

```
CREATE OR REPLACE VIEW gold.secured_product_sales_analysis AS
SELECT *
FROM gold.product_sales_analysis
WHERE
    CASE
        WHEN CURRENT_USER() = 'sales_user1@mukeshbackup09gmail.onmicrosoft.com'
        THEN category_name IN ('Bikes')
        WHEN CURRENT_USER() = 'marketing_user1@mukeshbackup09gmail.onmicrosoft.com'
        THEN category_name IN ('Clothing', 'Accessories')
        ELSE TRUE
    END;
```

1.2 RLS View for gold.customer_purchase_analysis

```
CREATE OR REPLACE VIEW gold.secured_customer_purchase_analysis AS
SELECT *
FROM gold.customer_purchase_analysis
WHERE
    CASE
        WHEN CURRENT_USER() = 'sales_user1@mukeshbackup09gmail.onmicrosoft.com'
        THEN region IN ('Germany')
        WHEN CURRENT_USER() = 'marketing_user1@mukeshbackup09gmail.onmicrosoft.com'
        THEN region IN ('Northwest')
        ELSE TRUE
    END;
```

1.3 RLS View for gold.monthly_sales_trends

```
CREATE OR REPLACE VIEW gold.secured_monthly_sales_trends AS
SELECT *
FROM gold.monthly_sales_trends
WHERE
    CASE
        WHEN CURRENT_USER() = 'sales_user1@mukeshbackup09gmail.onmicrosoft.com'
        THEN region IN ('Germany')
        WHEN CURRENT_USER() = 'marketing_user1@mukeshbackup09gmail.onmicrosoft.com'
        THEN region IN ('Northwest')
        ELSE TRUE
    END;
```

1.4 RLS View for gold.territory_performance

```
CREATE OR REPLACE VIEW gold.secured_territory_performance AS
SELECT *
FROM gold.territory_performance
WHERE
    CASE
        WHEN CURRENT_USER() = 'sales_user1@mukeshbackup09gmail.onmicrosoft.com'
        THEN region IN ('Germany')
        WHEN CURRENT_USER() = 'marketing_user1@mukeshbackup09gmail.onmicrosoft.com'
        THEN region IN ('Northwest')
        ELSE TRUE
    END;
```

Chapter 6

Dashboards

Dashboards were designed with role-specific access. Only relevant users were given access to their respective dashboards, ensuring secure and actionable insights.

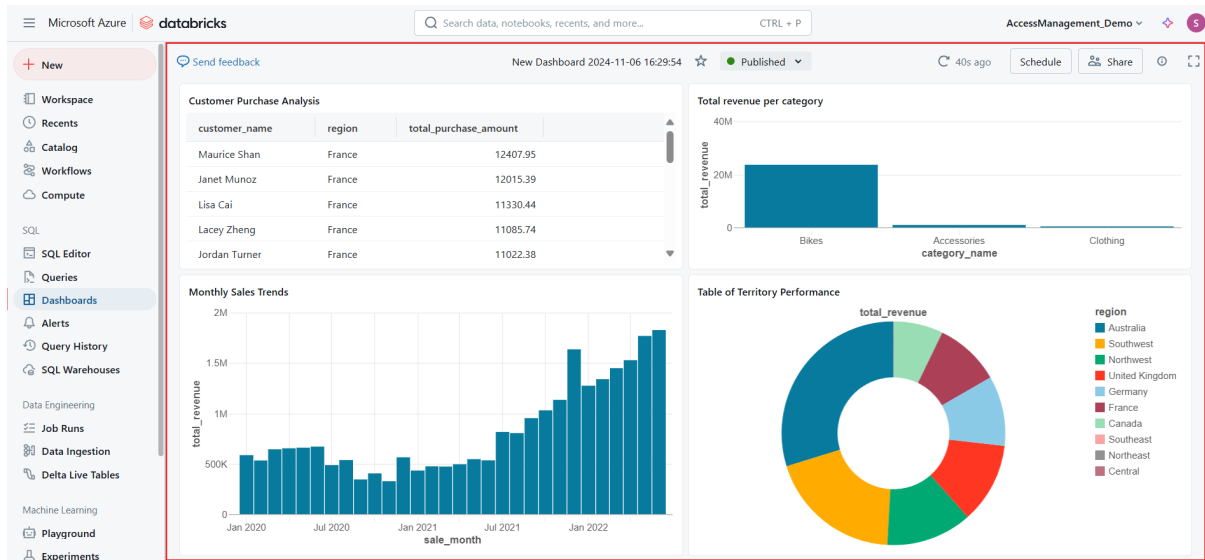


Figure 6.1: Role-based dashboards with restricted access.

Chapter 7

Data Cataloging with Unity Catalog

7.1 Centralized Metadata Repository

Unity Catalog was used to catalog datasets across all layers. The key features implemented include:

- **Data Lineage:** Enabled data lineage tracking for better traceability and governance.
- **Centralized Metadata Management:** Provided a single repository for managing metadata, improving discoverability and collaboration.

Data Lineage for accessmanagement_demo.gold.product_sales_analysis

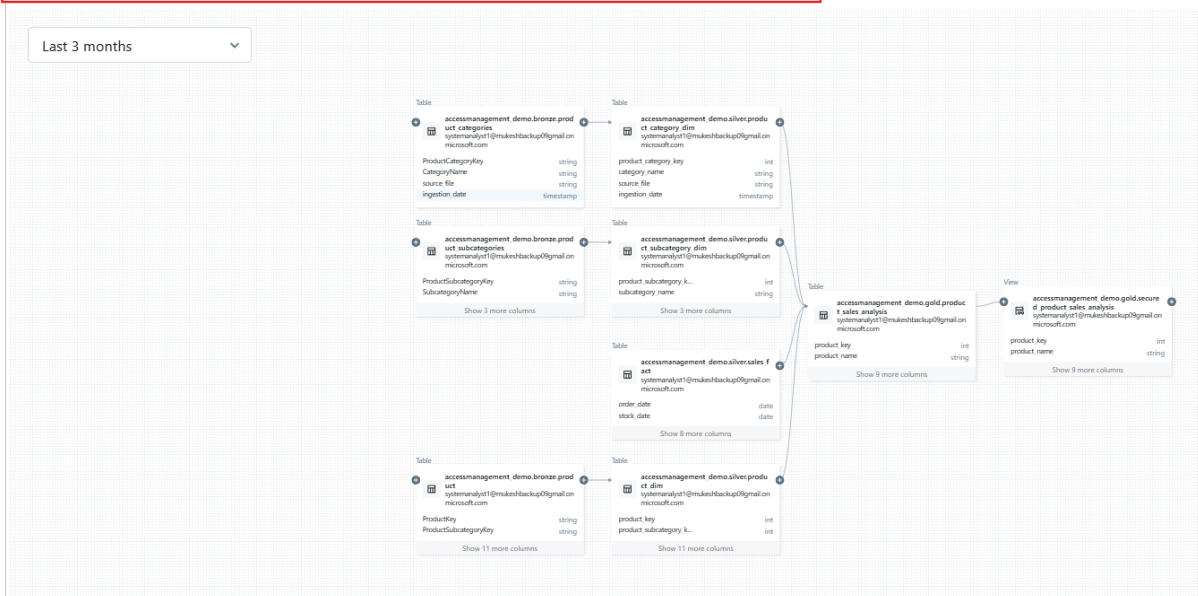


Figure 7.1: Data Lineage in Unity Catalog

7.2 Enhanced Data Governance

By implementing Unity Catalog, we achieved:

- **Improved Governance:** Enforced consistent policies for data access and usage.
- **Compliance:** Ensured that all datasets comply with organizational and regulatory requirements.

Chapter 8

Lessons Learned

The implementation of the Data Access Management system offered several critical insights:

- **Role-Based Design:** Designing the system around role-based access ensures secure data democratization while maintaining strict control over sensitive data.
- **SCIM Integration:** Automated user provisioning and role synchronization through SCIM significantly streamlined user management processes.
- **Unity Catalog Advantages:** The centralized governance offered by Unity Catalog enhanced operational efficiency by simplifying metadata management and improving data lineage tracking.
- **Challenges:**
 - Limited functionality due to subscription constraints.
 - Initial setup complexity with SCIM and Unity Catalog.

Chapter 9

Conclusion

This project successfully demonstrates the implementation of a secure and scalable Data Access Management System using Azure Databricks and Unity Catalog. The system effectively:

- Enforces role-based access control (RBAC) and row-level security (RLS) to protect sensitive data.
- Facilitates efficient data governance through Unity Catalog's centralized metadata repository and data lineage capabilities.
- Empowers stakeholders with secure access to actionable insights through role-specific dashboards.

The implementation highlights the critical role of data governance and access control in modern analytics platforms, ensuring both security and usability. Future work could involve expanding these capabilities with more advanced features like dynamic policy enforcement and integrating additional governance tools.

Chapter 10

Code File

For a detailed overview of the code and scripts used, refer to the following links:

- [Demo Code \(HTML\)](#) Open using a web browser
- [Demo Code \(IPYNB\)](#)

References

The following resources were consulted during the implementation of this project. These sources provided valuable insights and technical guidance for setting up the Data Access Management system, including user management, data governance, and role-based access control in Azure Databricks and Unity Catalog.

- **[Azure Databricks Getting Started Guide](#)**
This guide provides an overview of how to get started with Azure Databricks, including workspace setup, cluster creation, and data exploration.
- **[How to Create and Delete Users in Azure Entra ID](#)**
This documentation explains the process of managing users in Azure Entra ID, including creating, deleting, and assigning roles to users.
- **[Connecting Azure Storage to Databricks](#)**
A detailed guide on connecting Azure Data Lake Storage (ADLS) to Databricks for efficient data access and storage.
- **[SCIM Integration with Azure Active Directory](#)**
This resource explains how to configure SCIM integration between Azure Active Directory and Databricks for seamless user and group management.
- **[Unity Catalog Getting Started Guide](#)**
Provides a comprehensive introduction to Unity Catalog, including setup, data governance features, and best practices for managing metadata and access controls.